INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 161-165

DETECTING SPAM BY APPLYING MACHINE LEARNING APPROACH OVER EMAIL

Manish Kumar Sen[1], Pankaj Richhariya[2]

[1]Department of Computer Science, Bhopal Institute of Technology & Science, Bhopal, India.

ABSTRACT

These days, machine learning algorithms are effectively used to automatically screen spam emails. In this paper, we analyze several popular machine learning methods and explain their efficacy in detecting spam emails. Email is one of the most popular forms of communication since it is easily accessible, enables quick message exchanges, and has a low transmission cost. Email is the fastest and least expensive way to communicate. Spam in emails is one of the most challenging problems with email systems. Spam emails are unsolicited, unsolicited messages delivered for commercial, fraudulent, or other purposes that are not addressed to a specific recipient. Many methods of detection and filtering are used to keep the spam under control. The KNN algorithm, which is a content-based approach, is one of the most practical and straightforward methods. For the purpose of folder and subject classifications in this study, a sizable collection of personal emails was first employed. To make the next KNN algorithm more time-efficient, it is improved and expanded to I-KNN. Later, this improved algorithm is applied as a whole to produce superior Email Spam Detection results. Java is used for the implementation, and the parameters are calculation time and similarity score. Obtained result shows that the efficiency of proposed approach is better than traditional approach.

Keywords: Spam, KNN, N-Gram, E-Mail Classification, Machine Learning Algorithm.

1. INTRODUCTION

The present SPAM, or called garbage or illegal email is a long ways from 1937 and has nothing to do with tackling issues, and everything to do with made them. Spammers may contend that it's not so much an issue and that you can just erase what you don't need, yet that is credulous in the outrageous. Regardless of whether got as a private client, or as a business proficient, SPAM is a digital danger for various reasons:-

On the most fundamental dimension, getting a huge volume of SPAM messages squanders profitable transmission capacity and time. Private client are compelled to separately erase their undesirable message and overseers need to battle comparative issues yet on a far bigger scale trying to keep our organizations operational. This sat idle and exertion unavoidably prompts a misfortune in profitability as important assets are wastefully designated to none gainful endeavors. SPAM is a prime methods for exchanging electronic infections and malware diseases, regardless of whether purposely, or coincidentally as the immediate consequence of creating mass messages to and from an expansive number of beneficiaries. SPAM is likewise not just confined to mass showcasing plans. Perhaps the best issue is the disturbance factor. Email and the web can be viewed as one of humankind's most noteworthy innovations, yet we are in peril of wasting a standout amongst the most amazing assets of correspondence this planet has ever known. Maybe it is a human contrition to pulverize all that it has made, or perhaps it is only a similar old tricks that there have dependably been, just made progressively noticeable by a really worldwide correspondences medium.

1.1 Review Spam

Online product reviews have become an indispensable resource for users for their decision making while making online purchases. Product reviews provide information that impacts purchasing decisions to consumers, retailers, and manufacturers. Consumers make use of the reviews for not just a word of mouth information about any product, regarding product durability, quality, utility, etc. but also to give their own input regarding their experience to others. The rise in the number of E-commerce sites has lead to an increase in resources for gathering reviews of consumers about their product experiences. As anyone can write anything and get away with it, an increase in the number of Review Spams has been witnessed. There has been a growth in deceptive Review Spams - spurious reviews that have been fabricated to seem original [1]. These reviews produced by people who do not have personal experience on the subjects of the reviews are called spam, fake, deceptive or shill reviews. These spammers publish fictitious reviews in order to promote or demote a targeted product or a brand, convincing users whether to buy from a particular brand/store or not[2]. In the last few years, Review Spam Detection has gathered a lot of attention. Over the past few years, consumer review sites like Yelp.comhave been removing spurious reviews from their website using their own algorithms. Both supervised as well as unsupervised learning approaches have been used previously for filtering of Review Spams. For the purpose of training the features for machine learning approaches, linguistic and behavioural features have been used.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 161-165

There are two distinct types of deceptive review spams:

1. Hyper spam, in which fictitious positive reviews are rewarded to products to promote them

2. Defaming spam, where unreasonable negative reviews are given to the competing products to harm their reputations among the consumers [3] Specifically, the reviews that have been written either to popularize or benefit a brand or a product , therefore expressing a positive sentiment for a product, are called positive deceptive review spams. As opposed to that, reviews that intend to malign or defame a competing product expressing a negative sentiment towards the product, are called negative deceptive review spams[4].

1.2 Problem Definition

The accuracy, F-measure, Precision, and Recall parameters of the KNN algorithm were expanded by the authors in this study, and these parameters were later combined to provide superior Email Spam Detection results. Every time the training phase of the KNN algorithm is run, its internal centers and values are recalculated. KNN may thus make new computations and determine the closest match. I-KNN will become a time-efficient approach as a result of the authors' revised KNN algorithm, which only requires one computation. Additionally, I-KNN's single computation will increase accuracy as well.

2. LITERATURE SURVEY

Data mining techniques are used by R. Kishore Kumar, G. Poonkuzhali, and P. Sudhakar to analyze email spam classifiers. Spam dataset is studied using TANAGRA data mining tool in their paper, "Comparative Study on Email Spam Classifier using Data Mining Techniques," to discover the most effective classifier for email spam categorization. To extract the pertinent features, feature construction and feature selection are first performed. Then, different classifiers are applied to this dataset using different classification techniques, and each classifier is cross-validated. Based on the error rate, accuracy, and recall, the best classifier for email spam is found. [4].

User emails were categorized to prevent spam infiltration by Rafiqul Islam and Yang Xiang. They provide an effective and efficient email classification approach based on data filtering in their work, "Email Classification Using Data Reduction Method." In order to exclude the unnecessary data instances from the training model and subsequently categorize the test data, they have developed an original filtering strategy utilizing the instance selection method (ISM). The goal of ISM is to determine, with little information loss, which instances (examples, patterns) from email corpora should be chosen as representative of the full dataset. They tried several classification methods and employed the WEKA interface in our integrated classification model. Their empirical experiments demonstrate significant classification accuracy performance with a decrease in false positive occurrences. [5].

"Spam Detection Using Bayesian with Pattern Discovery" was a piece of work by Asmeeta Mali. In her study, she offers a useful method for enhancing the efficiency of utilizing and updating patterns found when looking for intriguing and pertinent material. With a high degree of term accuracy, we can identify spam emails from the email dataset using the Bayesian filtering method and an efficient pattern discovery technique. [6].

An picture spam detection method using detect spam terms is suggested by VandanaJaswal. In her research, a filtering technique called "Spam Detection System Using Hidden Markov Model" is utilized to find the phrases that make up spam images before using Hidden Markov Model spam filters to find all the spam photos. [7].

SaadatNazirova performed a piece titled "Survey on Spam Filtering Techniques" in 2011. This document provides an overview of the various email spam filtering techniques currently in use. Traditional and learning-based approaches are categorized, assessed, and compared. A few individual anti-spam products are evaluated and contrasted. The claim for a novel spam filtering technology is taken into consideration. [8].

3. PROPOSED FRAMEWORK

The k-closest neighbor (K-NN) classifier is considered an instance-based classifier, which means that preparation archives rather than a clear-cut classification depiction, such as the classification profiles used by other classifiers, are used for correlation. In that sense, there isn't really a stage of preparation. When another report has to be sorted, the k closest records (neighbors) are located, and if a sufficient amount of them have been assigned to a certain classification, the new archive is also placed in this classification, otherwise not. Additionally, conventional request methods can be used to discover the closest neighbors. We examine the class of the messages that are most similar to a message to determine if it is genuine or not. A persistent approach is the link between the vectors. The computation's likelihood of the k nearest neighbors is as follows:

Stage 1: Setting up

The preparatory messages should be saved.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 161-165

Sifting at stage 2

Pick from among the messages in the readiness set a message x's k nearest neighbors. Group the provided message as spam if there are other spam messages nearby. In general, consider it to be official mail. It is important to keep in mind that using an ordering technique to shorten the examination period results in an expansion of the model with a multidimensional nature O(m), where m is the prior measure. This approach is similarly recommended as a memory-based classifier since the majority of planning reference points are stored in memory. Another problem with the computation that is shown is that there is supposedly no parameter that we might adjust to reduce the number of false positives. By switching the arrangement guideline to the accompanying l/k rule, this problem is effectively resolved:

If l or more of message x's k nearest neighbors are spam, consider x to be spam and normally treat x as valid mail.

When it comes to arranging assignments, the k nearest neighbor rule has been widely used. It is also one of the very few universally consistent characterisation rules.

4. SIMULATION ENVIRONMENT

We are creating a program that offers two-factor authentication. JDK 1.8, a JAVA development kit, is what we're utilizing to create our program. A programming language is JAVA. It facilitates the developer's use of English-based commands while writing computer instructions. This kind of language is referred to as high-level language since it is stable and simple for a person to write. The format of the instruction is predetermined by a set of rules in JAVA. Its syntax is referred to as these rules. High-level instructions are converted into numerical codes that computers can comprehend and carry out once a program has been built. Enterprise software and web-based content. A software development kit (SDK) for creating Java programs is known as the JAVA development kit. The JDK is developed by Oracle INC Java soft division .

4.1 System Description

The method is built on a Java platform with 8 GB of RAM, 1 TB of HDD, and documents from the Pubnet dataset that contain various documents relevant to spam content. The analysis is carried out using an implementation based on the Swing & Chart API. The computation of the result and the parameter computation time are done in milliseconds, and the computation of the similarity measure is done in percent. The results computed using KNN, which is an improvement of KNN based with robin karp spamming technique, are shown below. The method determines its effectiveness by comparison with two factors.

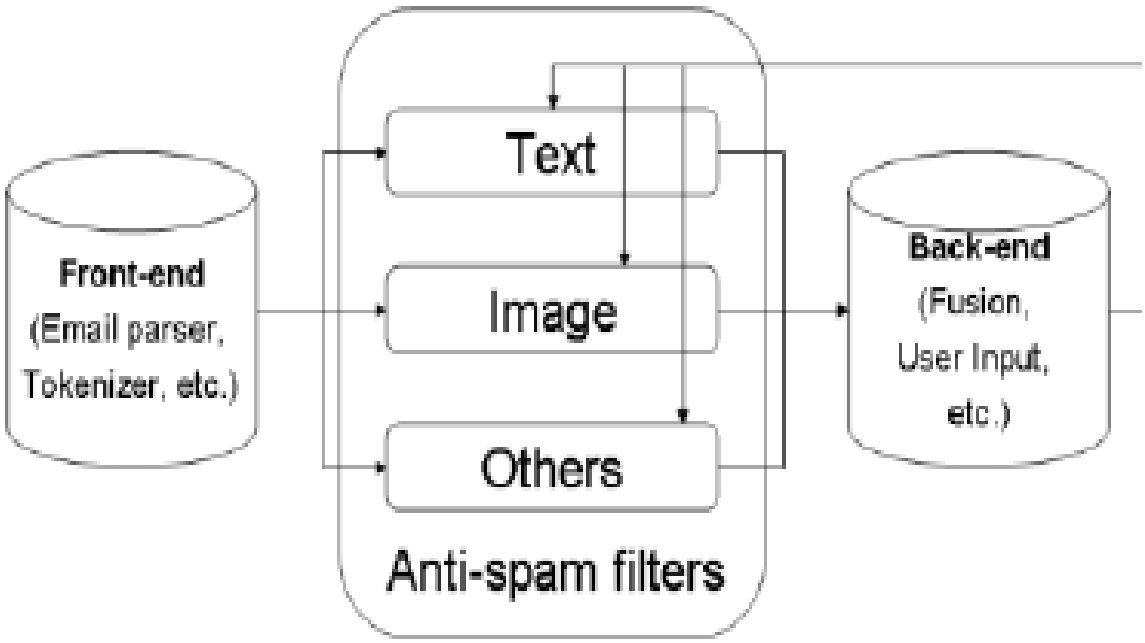

Figure 1: Block diagram.

4.2 Dataset

Recruitment Post - Access to a big dataset is necessary for the experimental setup and analysis stage of the process, thus we developed our own dataset that contains postings with dates for the analysis. Our dataset was designed to include a variety of messages, including ordinary posts (text from an everyday user) and posts containing sensitive material (text from a violent group). The dataset was then divided into two categories: Extreme Violation Detected---YES or NO.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 12, December 2023, pp : 161-165

4.3 Observation

The same dataset as previously described was used for the findings, and both methodologies were applied to it in order to obtain comparable results. For the analysis, we developed our own dataset that includes postings with dates. Our dataset was designed to include a variety of messages, including ordinary posts (text from an everyday user) and posts containing sensitive material (text from a violent group). The dataset was then divided into two categories: Extreme Violation Detected --- YES or NO…

Table 1: Comparison analysis obtained through Traditional spamming approach vs. KNN proposed scenario.

Algorithm	Computation time in msec.	Similarity measure
Traditional approach	86.72	50.00
KNN	68.72	64.32

The above table 1, discuss about the computation observed during the result computation.

Comparison analysis graphical representation:-

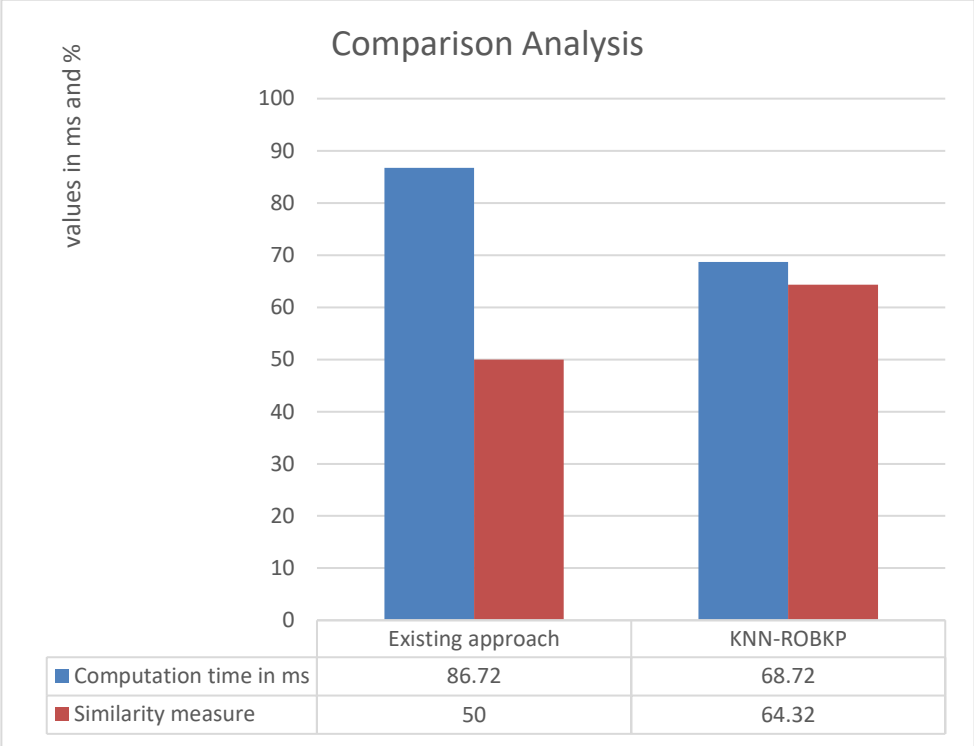

Figure 2: Comparison analysis graphical representation.

The graphical representation in figure 2 above, shows the comparison and efficiency of proposed algorithm using the given parameter. And hence the proposed algorithm is very useful while dealing with spamming analysis over the data using KNNalgorithm.

5. CONCLUSION

On the one hand, technology is giving us many benefits, but on the other, there is a drawback to email use: junk mail, which may be sent to anyone's email address by a third party and is regarded as spam or junk mail. Five machine learning methods for anti-spam filtering are suggested in this research. So, the study provides a summary of spa filtering methods.Although the term "spam" cannot be defined in a specific way, it may be said that spam communications are unsolicited. It causes a ton of issues for the ethical and economic sectors, which is why attempts have been made to define and outlaw spam through legislation. KNN-classifier based filtering is the method for anti-spam that is both highly advised and employed. Currently, a variety of artistic forms are used to organize emails according to various criteria in various regions. It frequently refers to the automated processing of incoming communications. However, this phrase is also used to refer to the human intervention in current messages as well as those that have already been received. As a result, the old N-gram strategy may be replaced with the proposed efficiency approach. Applying the suggested technique to real-time datasets can be used in subsequent research. IDS methods can also be used to improve the strategy.

6. REFERENCES

[1] Aladdin Knowledge Systems, "Anti-spam white paper, www.csisoft.com/security/aladdin/esafe_antispam", Retrieved December 28, 2011.

[2] F. Smadja, H. Tumblin, "Automatic spam detection as a text classification task", in: Proc. of Workshop on Operational Text Classification Systems, 2016.

[3] Ann Nosseir , Khaled Nagati and Islam Taj-Eddin, "Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013 ISSN (Print): 1694- 0814 | ISSN (Online): 1694-0784.

[4] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar," Comparative Study on Email Spam Classifier using Data Mining Techniques", Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, IMEC2012, March 14- 16,2012, Hong Kong, ISBN: 977-988-19251-1-4.

[5] Rafiqul Islam and Yang Xiang, member IEEE, "Email Classification Using Data Reduction Method" created June 16, 2010.

[6] Asmeeta Mali, "Spam Detection Using Baysian with Pattren Discovery", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-3, July 2013.

[7] VandanaJaswal, " Spam Detection System Using Hidden Markov Model", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013 ISSN: 2277 128X.

[8] [8] SaadatNazirova, "Survey on Spam Filtering Techniques", Communications and Network, 2011, 3, 153 160, doi:10.42 36/cn.2011.33019 Published Online August 2011 (http: //www.SciRP.org /journal/cn).

[9] [9] Neha Singh,"Dendritic Cell Algorithm and Dempster Belief Theory Using Improved Intrusion Detection System ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013 ISSN: 2277 128X.

[10] Julie Greensmith, "The Dendritic Cell Algorithm", Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy October 2007.

[11] R. O. Duda, P. E. Hart, and D. G. Stork.Pattern Classification.Wiley, New York, 2.edition, 2001.

[12] R. O. Duda, P. E. Hart, and D. G. Stork.Pattern Classification.Wiley, New York, 2.edition, 2001.

[13] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. Advances in knowledge discoveryand data mining.American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[14] K. Feng, J. Gao, K. Feng, L. Liu, and Y. Li. Active and passive nearest neighbor algorithm: A newlydevelopedsupervised classifier. In D.-S.Huang, Y. Gan, P. Gupta, and M. Gromiha, editors, AdvancedIntelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, volume 6839 ofLecture Notes in Computer Science, pages 189–196. Springer Berlin Heidelberg, 2012.

[15] S. Garcia and F. Herrera. Evolutionary undersampling for classification with imbalanced datasets: proposalsand taxonomy. Evolutionary Computation, 17(3):275–306, 2009.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data miningsoftware: an update. SIGKDD Explor.Newsl., 11:10–18, November 2009.

[17] B. Hamers, J. A. K. Suykens, and B. D. Moor. Compactly Supported RBF Kernels for Sparsifying the GramMatrix in LS-SVM Regression Models. 2002.

[18] J. Han and M. Kamber.Data Mining: Concepts and Techniques, volume 54. Morgan Kaufmann, 2006.

[19] S. Haykin. Neural Networks: A Comprehensive Foundation. Prentice-Hall, 1999.

[20] D. Hosmer and S. Lemeshow. Applied Logistic Regression. Wiley Series in Probability and Statistics: Textsand References Section. John Wiley & Sons, 2000.

[21] J. P. Hwang, S. Park, and E. Kim.A new weighted approach to imbalanced data classification problem viasupport vector machine with quadratic cost function.Expert Syst. Appl., 38:8580–8585, July 2011.

[22] L. Hyafil and R. L. Rivest. Constructing Optimal Binary Decision Trees is NP-Complete. InformationProcessing Letters, 5:15–17, 1976.

[23] M. Z. Jahromi, E. Parvinnia, and R. John. A method of learning weighted similarity function to improvethe performance of nearest neighbor. Inf. Sci., 179:2964–2973, August 2009.

[24] T. M. Khoshgoftaar, S. Zhong, and V. Joshi. Enhancing software quality estimation using ensembleclassifierbased noise filtering. Intell. Data Anal., 9(1):3–27, Jan. 2005.

[25] E. Kriminger, J. Principe, and C. Lakshminarayan.Nearest neighbor distributions for imbalanced classification.In Neural Networks (IJCNN), The 2012 International Joint Conference on, pages 1 –5, june 2012.